International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 3 Issue 5, August 2019 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

A Hybrid Apporach of Classification Techniques for Predicting Diabetes using Feature Selection

S. Jaya Mala

Research Scholar, Bishop Heber College, Trichy, Tamil Nadu, India

How to cite this paper: S. Jaya Mala "A Hybrid Apporach of Classification Techniques for Predicting Diabetes using Feature Selection" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-3 | Issue-5, August 2019, pp.2506-2510, **https://doi.org/10.31142/ijtsrd27991**

IJTSRD27991

ABSTRACT

Diabetes is predicted by classification technique. The data mining tool WEKA has been developed for implementing Support Vector Machine (SVM) classifier. Proposed work is framed with a specific end goal to improve the execution of models. For improving the classification accuracy Support Vector Machine is combined with Feature Selection and percentage Split. Trial results demonstrated a serious change over in the current Support Vector Machine classifier. This approach enhances the classification accuracy and reduces computational time.

KEYWORDS: *Data Mining, Diabetes, Classification, SVM, J48, Naïve Bayes*

1. INTRODUCTION

Classification strategies are broadly used in the medical field for classifying data into different classes according to some constrains comparatively an individual classifier. Diabetes is an illness which affects the ability of the body in producing the hormone insulin, which in turn makes the metabolism of carbohydrate abnormal and raise the levels of glucose in the blood. In Diabetes a person generally suffers from high blood sugar. Intensify thirst, Intensify hunger and Frequent urination are some of the symptoms caused due to high blood sugar. Many complications occur if dia-betes remains untreated. Some of the severe complications include diabetic ketoacidosis and Nonketotic hyperosmolar coma [1]. Diabetes is examined as a major health issues during which the measure of sugar substance cannot be controlled.

Diabetes is not only pretentious by various factors like height, weight, hereditary factor and insulin but the major reason considered is sugar concentration among all factors. The early credentials is the only remedy to stay away from the complications. Many researchers are conducting various experiments for diagnosing the diseases using various classification algorithms of machine learning approaches like J48, SVM, Naive Bayes, Decision Tree, Decision Table etc. as researches have proved that machine-learning algorithms works better in diagnosing different diseases. Data Mining and Machine learning algorithms gain its strength due to the capability of managing a large amount of data to combine data from several different sources and integrating the background information in the study [2]. This research work concentrations on pregnant women suffering from diabetes. In this work, Naive Bayes, SMO and J48 machine learning classification algorithms are used and calculated on the PIDD dataset to find the prediction of diabetes in a patient.

2. RELATED WORK

Deepti Sisodia et al [3] proposed this paper one of the significant genuine medical issues is the detection of diabetes at its beginning time. In this examination, precise endeavors are made in structuring a framework which results in the prediction of ailment like diabetes. During this work, three machine learning characterization calculations are considered and assessed on different measures. Tests are performed on Pima Indians Diabetes Database. Experimental results decide the sufficiency of the planned framework with an accomplished precision of 76.30 % utilizing the Naive Bayes order algorithm. In future, the structured framework with the utilized machine learning techniques order calculations can be utilized to anticipate or analyze different illnesses. The work can be expanded and improved for the mechanization of diabetes examination including some other Machine learning algorithms.

Fikirte Girma Woldemichael et al.[4] designed this paper data mining is the way toward extracting helpful and previously obscure examples from huge database or Data ware house. Presently data mining assumes significant job in numerous parts a portion of this are in wellbeing division, Bank, financial part, training segment and so forth. Various examines have been done about an expectation of diabetes utilizing distinctive calculation. This investigation is proposed to anticipate diabetes mellitus utilizing Back spread calculation. Also, other characterization calculation like J48, Naive bayes and support vector machine calculation were utilized to anticipate diabetes.5-cross fold procedure and huge worth learning rate were utilized to improve the precision of back engendering calculation. PIMA Indian dataset were utilized as information informational collection for foreseeing diabetes. R programming language was accustomed to actualizing this examination. The investigation created [8-6-1] neural system design to foresee diabetes. Back propagation calculation gives 83.11% prescient exactness, these outcomes were gotten with less number of cycles and it demonstrates improvement from past paper. As per the consequence of this work exactness of Back propagation in prediction of diabetes is superior to SVM, J48 and Naive Bayes calculation.

Aiswarya Iyer et al [4] worked this paper the programmed diagnosis of diabetes is a significant real world restorative issue. Detection of diabetes in its beginning times is the key for treatment. This paper indicates how decision Trees and Naive Bayes are utilized to display real finding of diabetes for nearby and orderly treatment, alongside showing related work in the field. Exploratory outcomes demonstrate the adequacy of the proposed model. The exhibition of the systems was examined for the diabetes finding issue. experimental results exhibit the sufficiency of the proposed model. In future it is wanted to accumulate the data from various regions over the world and make a progressively exact and general insightful model for diabetes end. Future examination will in like manner concentrate on get-together data from a later timeframe and find new potential prognostic components to be fused. The work can be expanded and improved for the robotization of diabetes examination.

3. PROPOSED WORK AND METHODOLOGY

The proposed work implemented in Weka tool. Pima Indian dataset were used for implementing this study.

A. Algorithm

The following classification algorithms are used in proposed methods:

- J48 Algorithm
- Naïve Bayes classifier
- SMO

J48 algorithm:

J48 is supervised learning algorithm. It is an extension of ID3 algorithm, it has the additional features like handling missing value, decision trees pruning, continuous attribute value ranges, derivation of rules, etc. j48 form classification using by training the model and corroborate the model using test cases.

Similarly on the bases of the arrangement events the classes for the as of recently created models are being found. This figuring produces the benchmarks for the desire for the goal variable. With the help of tree course of action count the fundamental flow of the data is viably understandable. In the WEKA data mining furnishes, J48 is an open source Java utilization of the C4.5 estimation. The WEKA apparatus outfits different choices related with tree pruning. In the event that there ought to be an event of potential over fitting pruning can be used as a gadget for précising. In various estimations the course of action is performed recursively till every single leaf is unadulterated, that is the arrangement of the information performed to be as faultless as would be conceivable. This computation it delivers the principles from which explicit character of that data is made. The objective is intelligently speculation of a decision tree until the moment that it grabs equalization of adaptability and exactness.

Naïve Bayes Classifier:

Naive Bayes is a classification technique with a notion which defines all features are independent and unrelated to each other. It defines that status of a specific feature in a class does not affect the status of another feature. Since it is based on conditional probability it is considered as a powerful algorithm employed for classification purpose. It works well for the data with imbalancing problems and missing values. Naive Bayes [7] is a machine learning classifier which employs the Bayes Theorem. Using Bayes theorem posterior probability $P(C|X)$ can be calculated from P(C),P(X) and $P(X|C)$ [6]. Therefore,

$$P(C|X) = (P(X|C)\ P(C))/P(X)$$

Where, $P(C|X)$ = target class's posterior probability. $P(X|C)$ = predictor class's probability. P(C) = class C's probability being true. P(X) = predictor's prior probability.

SVM (Support Vector Machine)

Support Vector Machines are a decently current sort of learning algorithms, at first exhibited. Regularly, SVM go for pointed for the hyper plane that most phenomenal segregates the classes of data. SVMs have affirmed the limit not only to explicitly detach substances into change classes, yet although perceive case whose set up gathering isn't maintained by data. Despite the way that SVM are almost hard describe movement of getting ready instances of each class. SVM can be simply connected with perform numerical checks. Two such enlargements, the first is to expand SVM to execute relapse investigation, where the goal is to convey an immediate limit that can really correct that goal limit. An extra development is to make sense of how to rank parts as opposed to making a request for solitary segments. Situating can be diminished to taking a gander at sets of model and making a +1 assess if the consolidate is in the correct situating demand in extension to −1 something different.

B. Proposed work:

This study, data were collected from PIMA Indian dataset. Back propagation algorithms were implemented to predict diabetes diseases. This study implemented in Weka tool

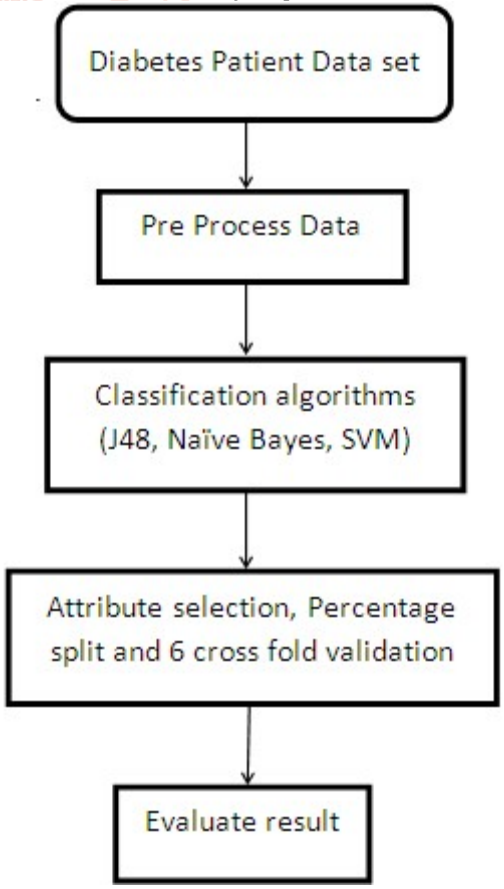

Fig 1: Frame work for proposed algorithm

C. Dataset Description

This study use PIMA Indian dataset which is downloaded from UCI machine learning repository. PIMA Indian dataset

consists of 768 instances with having 8 attributes and one class attribute with having two class labels. Attribute description are listed below in the table 1.

S. No	Name	Description
1	Preg	No of times pregnant
2	Plas	Plasma glucose concentration a 2 hours in a oral glucose tolerance test
3	Pres	Diastolic blood pressure (mm hg)
4	Skin	Triceps skin fold thickness (mm)
5	Insu	2 hours serum insulin (mm u/ml)
6	Mass	Body mass index (weight in kg/ (height in m)^2)
7	Pedi	Diabetes Pedigree Function
8	Age	Age (years)
9	Class	Class Variable (0 or 1)

Table1: Attributes Of Diabetes Dataset

D. Data preprocessing

Data preprocessing step is one important step in knowledge discovery process (KDD). Most health care data contain missing value, noisy and inconsistency data. Pima Indian data set has missing value and inconsistency data. Missing value are recognized then replaced or moved by attribute meanvalue. Inconsistence data recognized and removed manually. The dataset doesn't consist of noisy data.

E. Feature selection

Feature selection is process of selecting useful feature from dataset. Chi-square test feature selection technique was used for selecting importance feature to predict diabetes disease from Pima Indian dataset. From 8 attribute chi-square test feature selection technique selected that glu ,Age, bmi,serum,Npreg ,skin are the most important attribute for predicting diabetes disease. The importance of each attribute for predicting diabetes diseases shown on the fig below:[8]

	attr_importance
Npreg	0.2359031
glu	0.5028310
bp	0.1647848
skin	0.2176562
serum	0.2447105
bmi	0.2953318
Ped	0.1690577
age	0.3089785

Fig.2. Importance of attributes

4. RESULT AND DISCUSSION

PIMA Indian dataset were used to examine for this study. The Dataset have 8 independent attribute and one dependent attribute class, for conducting this work those attribute were trained to predict diabetes diseases. In addition, these works identify the most important attribute that contributes for prediction of diabetes. This Study were implement by Weka tool.J48, naïve Bayes and Support vector machine algorithm were used to prediction diabetes. Cross validation technique was used training data's and evaluating analytical performance of the model. Confusion matrixes were used to visualize and to measure performance model by using accuracy, sensitivity and specificity of the algorithm.

A. Accuracy measures

Naive Bayes, SVM and J48 algorithms are used in this research work. Experiments are performed using internal cross-validation 10-folds and percentage split 66%. Accuracy, F-Measure, Recall, Precision and ROC (Receiver Operating Curve) measures are used for the classification of this work. Table-2 defines accuracy measures below:

Table2. Accuracy Measures

S. No	Measures	Definition	Formulas
1	Accuracy(A)	Accuracy determines the accuracy of the algorithm in predicting instances.	A=(TP+TN) / (Total no of samples)
2	Precision(P)	Classifier accuracy is measured by Precision.	P = TP / (TP+ FP)
3	Recall	To measure the classifier completeness or sensitivity, Recall is used.	R =TP / (TP+FN)
4	F-measure	F-Measure is the weighted average of precision and recall.	F=2*(P*R) / (P+R)
5	ROC	ROC (Receiver Operating Curve) curves are used to compare the usefulness of tests.	

Table3. Comparative Performance of Classification Algorithms on Various Measures

Classification Algorithms	Precision	Recall	Accuracy	F-Measure	ROC
Naïve Bayes	0.769	0.763	76.30	0.761	0.812
J48	0.735	0.738	73.28	0.736	0.753
SVM	0.773	0.769	77.03	0.763	0.725

Corresponding classifiers performance of classification algorithms have 10 fold cross validation and percentage split is 66% over Accuracy, Precision, F-measure, Recall and ROC values are listed in Table-3 Where, TP defines True Positive, TN defines True Negative, FP defines False positive, FN defines False Negative. The corresponding classifiers performance on the basis of Accuracy, Precision, F-measure, Recall and ROC values are listed in Table-3.The below figure 3 show that the graphical representation of classification of algorithms

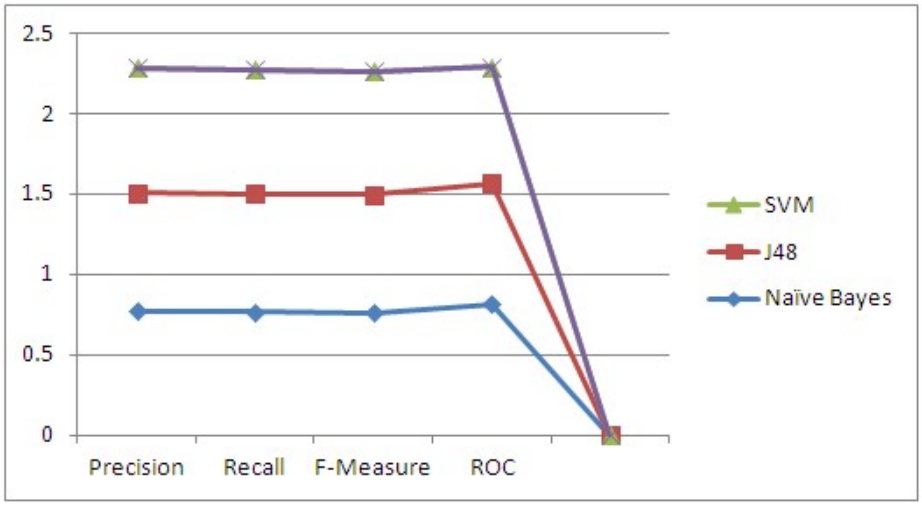

Fig.3 Classifiers performance of various measures

Table4: Comparative Performances of classification algorithms with attribute selection

Classification Algorithms	Precision	Recall	F-Measure	ROC
Naïve Bayes	0.771	0.776	0.771	0.827
J48	0.742	0.749	0.743	0.791
SVM	0.774	0.773	0.765	0.717

The above table shows that the corresponding classifiers performance of classification algorithms have attribute selection of pre- processed data, 6 cross fold validation and percentage split was 97% its show more accuracy.

Table-4 represents different performance values of all classification algorithms calculated on various measures. From Table-4 it is analyzed that Support Vector Machine (SVM) showing the maximum accuracy. So the SVM classifier can predict the chances of diabetes with more accuracy as compared to other classifiers. Performances of all classifier are based on various measures. The below figure 4 show that the graphical Representation of classification algorithms with attribute selection.

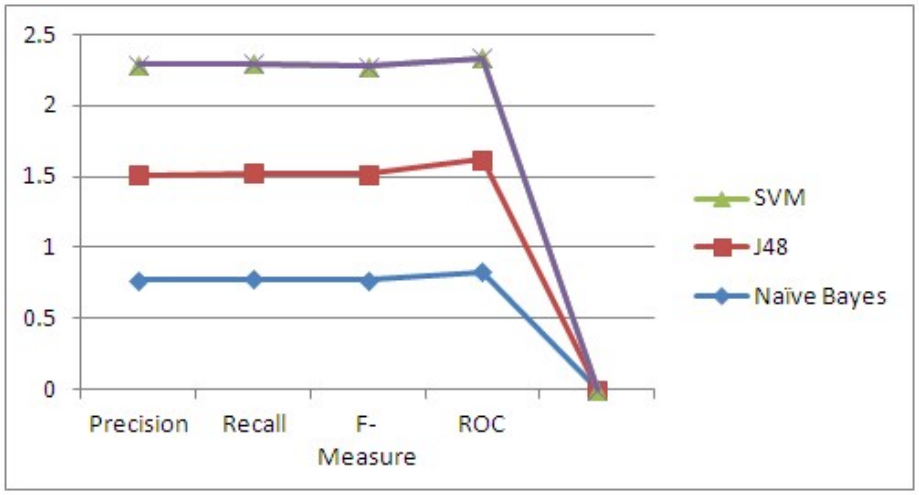

Fig4.Classifiers performance of various measures using attribute selection

Table 5: Classifier's Performance on The Basis of Classified Instances

Total No Of Instances	Classification Algorithms	Correctly classified instances	Incorrectly classified instances	Accuracy
768	Naïve Bayes	596	172	77.60
	J48	575	193	74.90
	SVM	594	174	78.30

Table-5 determines classifier performance on the basis of classified instances. According to these classified instances, accuracy is calculated and analyzed. Performance of individual algorithm is evaluated on the basis of Correctly Classified Instances and Incorrectly Classified Instances out of a total number of instances. Figure-4 shows the graphical performance of all classification algorithms on the basis of classified instances. From Table-4and Table-5 we can conclude that Support Vector Machine (SVM) classification algorithm outperforms comparatively other algorithms. So Support Vector Machine (SVM) is considered as the best supervised machine learning method of this experiment because it gives higher accuracy in respective to other classification algorithms with an accuracy of 78.30 %.

5. CONCLUSION

One of the most important real-world health care problems is the detection of diabetes at its early stage. In this study, efficient efforts are made in designing a system which results in the prediction of disease like diabetes. During this effort, three machine learning classification algorithms are studied and evaluated on various procedures. Experiments are performed on Pima Indians Diabetes Database. Experimental results determine the acceptability of the designed system with an achieved accuracy of 78.30 % using the Support Vector Machine (SVM) with Attribute selection, Cross fold Validation and Percentage split. In future, the designed system with the used machine learning classification algorithms can be used to predict or diagnose other diseases. The work can be extended and improved for the automation of diabetes analysis including some other machine learning algorithms.

6. REFERENCE

[1] Kumar, D. A., Govindasamy, R., 2015. Performance and Evaluation of Classification Data Mining Techniques in Diabetes. International Journal of Computer Science and Information Technologies, 6, 1312–1319.

[2] Fatima, M., Pasha, M., 2017. Survey of Machine Learning Algorithms for Disease Diagnostic. Journal of Intelligent Learning Systems and Applications 09, 1–16. doi:10.4236/jilsa.2017.91001.

[3] Deepti Sisodia, Dilip Singh Sisodia "Prediction of Diabetes using Classification Algorithms" International Conference on Computational Intelligence and Data Science (ICCIDS 2018) Procedia Computer Science 132 (2018) 1578–1585 (https://creativecommons.org/licenses/by-nc-nd/3.0/)

[4] Fikirte Girma Woldemichael, Sumitra Menaria" Prediction of Diabetes using Data Mining Techniques" Proceedings of the 2nd International Conference on Trends in Electronics and Informatics (ICOEI 2018 (IEEE Conference Record: # 42666; IEEE Xplore ISBN:978-1-5386-3570-4

[5] Aiswarya Iyer, S. Jeyalatha and Ronak Sumbaly "Diagnosis Of Diabetes Using Classification Mining Techniques" International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1, January 2015.

[6] Ray, S., 2017. 6 Easy Steps to Learn Naïve Bayes Algorithm (with code in Python).

[7] Rish, I., 2001. An empirical study of the naïveBayes classifier, in: IJCAI2001 workshop on empirical methods in artificial intelligence, IBM. pp. 41–46.

[8] Fikirte Girma Woldemichael, Sumitra Menaria," Prediction of Diabetes using Data Mining Techniques" Proceedings of the 2nd International Conference on Trends in Electronics and Informatics (ICOEI 2018) IEEE Conference Record: # 42666; IEEE Explore ISBN:978-1-5386-3570-4.